November 1996

LIDS-P-2372

Research Supported By:

AFOSR-F496-20-93-1-0604
ARO DAAL03-92-G-0015

A CANONICAL CORRELATIONS APPROACH TO MULTISCALE STOCHASTIC REALIZATION

William W. Irving Alan S. Willsky

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **NOV 1996**	2. REPORT TYPE	3. DATES COVERED **00-11-1996 to 00-11-1996**

4. TITLE AND SUBTITLE **A Canonical Correlations Approach to Multiscale Stochastic Realization**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Laboratory for Information and Decision Systems,77 Massachusetts Avenue,Cambridge,MA,02139-4307**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**40**	

LIDS-P-2372

A CANONICAL CORRELATIONS APPROACH TO MULTISCALE STOCHASTIC REALIZATION

William W. Irving Alan S. Willsky

Abstract

We develop a realization theory for a class of multiscale stochastic processes having white-noise driven, scale-recursive dynamics that are indexed by the nodes of a tree. Given the correlation structure of a 1-D or 2-D random process, our methods provide a systematic way to realize the given correlation as the finest scale of a multiscale process. Motivated by Akaike's use of canonical correlation analysis to develop both exact and reduced-order model for time-series, we too harness this tool to develop multiscale models. We apply our realization scheme to build reduced-order multiscale models for two applications, namely linear least-squares estimation and generation of random-field sample paths. For the numerical examples considered, least-squares estimates are obtained having nearly optimal mean-square errors, even with multiscale models of low order. Although both field estimates and field sample paths exhibit a visually distracting blockiness, this blockiness is not an important issue in many applications. For such applications, our approach to multiscale stochastic realization holds promise as a valuable, general tool.

*The work of William Irving and Alan Willsky was supported by the Air Force Office of Scientific Research, under grant AFOSR-F496-20-93-1-0604, the Army Research Office, under grant ARO DAAL03-92-G-0015.

1 Introduction

A class of stochastic processes indexed by the nodes of a tree was introduced in [4]. These processes have white-noise driven, scale-recursive dynamics, directly analogous to the time-recursive dynamics of Gauss-Markov time-series models. Experimental and theoretical results have demonstrated that this class of processes is quite rich statistically; the self-similarity of fractional Brownian motion can be represented [4], as can any given 1-D wide-sense (WS) reciprocal process or 2-D Markov random field (WSMRF) [12].[1] Complementing this statistical richness are the efficient image processing algorithms to which multiscale models lead. For example, a scale-recursive algorithm has been developed that directly generalizes the Kalman filter and related Rauch-Tung-Striebel (RTS) smoother [4]. This algorithm incorporates noisy measurements of a given multiscale process to calculate both smoothed estimates *and* associated error covariances. Another algorithm has been developed for likelihood calculation [11]. In contrast to traditional 2-D optimal estimation formulations based on Markov random fields, which have a per-pixel computational complexity that typically grows with image size. these multiscale algorithms have a per-pixel complexity independent of image size. Substantial computational savings can thus result, as evidenced by the work in [13] on calculating optical flow and the work in [7] on interpolation of sea level variations in the North Pacific Ocean from satellite measurements.

Just as Kalman filtering requires the prior specification of a state-space model, so does multiscale statistical processing. In this paper, we develop a general approach for building multiscale models. Given the correlation structure of a 1-D or 2-D random process[2] our methods provide a systematic way to realize the given correlation as the finest scale of a multiscale process. Because there is typically a conflict between model complexity and accuracy, we mainly focus on the case where a constraint is imposed on the allowed model state dimension; the objective then is to build a model whose finest-scale correlation structure best matches the desired correlation, subject to the

[1] The definition and properties of wide-sense reciprocal processes and WSMRFs are nicely summarized in [5].

[2] The terminology 1-D, 2-D or multidimensional random process is used here to indicate that the dimension of the independent variable of the process is 1-D, 2-D, or multidimensional.

dimension constraint. In general, our focus on realizing *finest*-scale statistics is motivated by the not insignificant class of applications in which the finest-scale is really the only one of interest. For instance, in many de-noising applications, the finest-scale process is a pixel-by-pixel representation of the image, the measurements are noisy observations of each pixel, and the objective is to estimate the value of each image pixel. For such problems, the multiscale framework provides an efficient statistical approach for obtaining estimates and error covariances, even though every other aspect of these problems involves only the finest scale.

There is a close relationship between the multiscale stochastic realization problem and its more traditional, time-series counterpart. This relationship can be made clear once the Markov property of multiscale processes is noted. To describe the Markov property of multiscale processes, we first observe that in a q-th order tree each node has q children and a single parent, and hence partitions the remaining nodes into $(q+1)$ subtrees, one associated with each of these child and parent nodes. (A second-order tree, which is often used to index multiscale representations of time series, is illustrated in Figure 1.) Now, the Markov property states that if $x(s)$ is the value of the state at node s, then conditioned on $x(s)$ the states in the corresponding $(q+1)$ subtrees of nodes extending away from s are uncorrelated. The connection to the time-series realization problem is that in both contexts, the role of state information is to provide an information interface among subsets of the process. This interface must store just enough process information to make the corresponding process subsets conditionally uncorrelated. In the time-series case, this interface is between two subsets of the process (i.e., the past and the future), while in the multiscale case, the interface is among *multiple* (i.e., $(q+1)$) subsets of the process.

We exploit the connection between the time-series and multiscale realization problems by adapting to the multiscale context work done in [1] and [2] on reduced-order time-series modeling. The work in [1] and [2] addresses two issues. First, for exact realizations, a method is devised for finding the minimal dimension and corresponding information content of the state. Second, for reduced-order, approximate realizations, a method is devised for measuring the relative importance of the components of the information interface provided by the state, so that a decision can be made about which components to discard in a reduced-order realization. The latter is accomplished using a classical tool from multivariate statistics, namely *canonical correlation analysis* [8]. We decompose our multiscale problem of decorrelating jointly $(q+1)$ process subsets into a collection of q problems of decorrelating pairs of process subsets. We then demonstrate that with respect to a particular decorrelation metric, canonical correlation analysis can in principle be used to solve optimally each of the pairwise decorrelation problems. Furthermore, these pairwise solutions can be concatenated to yield a sub-optimal solution to the multi-way decorrelation problem. The solution to this decorrelation problem leads readily to values for all the multiscale model parameters.

We apply our realization scheme to build reduced-order multiscale models for two applications, namely linear least-squares estimation and generation of random-field sample paths. For the numerical examples considered, we obtain least-squares estimates having mean-square errors that are nearly optimal, even with multiscale models of very low order. Although both field estimates and field sample paths exhibit a visually distracting blockiness, this blockiness is not really an issue in many applications. For such applications, our approach to multiscale realization holds promise as a valuable, general tool.

This paper is organized in the following way. In Section 2, the multiscale framework is more formally introduced, and a measure of decorrelation is defined. In Section 3, the modeling problem is precisely formulated, and a solution is overviewed for the case that full-order, exact models are sought. In Section 4, the solution to the modeling problem is presented for the more challenging case that reduced-order, approximate models are sought. In Section 5, two numerical examples are presented, and finally in Section 6, a summary is provided, together with suggestions for future work. Details of the proofs are relegated to appendices at the end of the paper.

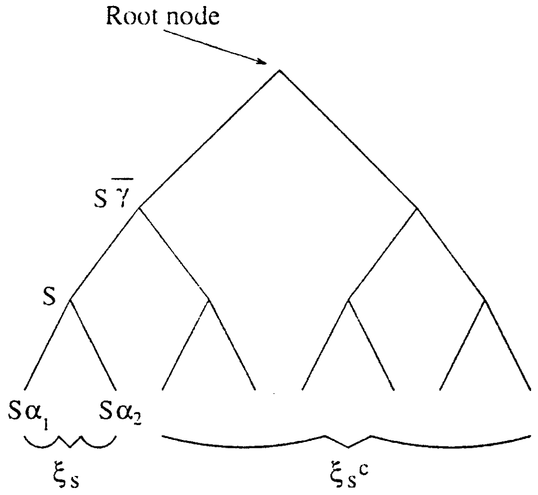


Figure 1: The first four levels of a 2-nd order tree are shown. The parent of node s is denoted by $s\bar{\gamma}$ and the two offspring are denoted by $s\alpha_1$ and $s\alpha_2$. The random vectors ξ_s and ξ_{s^c} contain, respectively, the finest-scale state information that does and does not descend from the node s.

2 Preliminaries

2.1 State-Space Models on q-th Order Trees

The models introduced in [4, 13] describe multiscale stochastic processes indexed by nodes on a *tree.* A q^{th}-order tree is a pyramidal structure of nodes connected such that each node has q offspring nodes. We associate with each node s a vector-valued state $x(s)$, where in general, the q^m state vectors at the m-th level of the tree (for $m = 0, 1, \ldots$) can be interpreted as information about the m-th scale of the process. In keeping with the conventions established in [4, 13], we define an upward (fine-to-coarse) shift operator $\bar{\gamma}$ such that $s\bar{\gamma}$ is the *parent* of node s, and a corresponding set of downward (coarse-to-fine) shift operators α_i, $i = 1, 2, \ldots, q$, such that the q offspring of node s are given by $s\alpha_1$, $s\alpha_2, \ldots, s\alpha_q$. Figure 1 depicts an example of the relative locations of $s, s\bar{\gamma}$, and $s\alpha_1, s\alpha_2$ in a second order tree.

The dynamics implicitly providing a statistical characterization of $x(s)$ have the form of an autoregression in scale:

$$x(s) \;=\; A(s)x(s\bar{\gamma}) \;+\; w(s). \tag{1}$$

This regression is initialized at the root node, $s = 0$, with a state variable $x(0)$ having mean zero and covariance $P(0)$. The term $w(s)$ represents white driving noise, uncorrelated across scale and space, and also uncorrelated with the initial condition $x(0)$; this noise is assumed to have mean zero and covariance $Q(s)$. Since $x(0)$ and $w(s)$ are zero-mean, it follows that $x(s)$ is a zero-mean random process[3]. Furthermore, since the driving noise $w(s)$ is white, the correlation structure of the process $x(s)$ is characterized completely by $P(0)$ and the autoregression parameters $A(s)$ and $Q(s)$ for all nodes $s \neq 0$.

The statistical structure of multiscale processes can be exploited to yield an extremely efficient algorithm for estimating $x(\cdot)$, based upon noisy observations $y(\cdot)$. These observations take the form

$$y(s) \quad = \quad C(s)x(s) + v(s),$$

where the noise $v(s)$ is white, has covariance $R(s)$, and is uncorrelated with $x(\cdot)$ at all nodes on the tree. Just like the Kalman filter and the RTS smoother, this estimation algorithm has a recursive structure, and yields both state estimates and associated error covariances. For a multiscale process having states of dimension λ and indexed on a tree with N nodes, the number of required computations is $\mathcal{O}(N\lambda^3)$. Thus, the algorithm is quite efficient, particularly when the dimension of the state vectors is low.

2.2 Markov Property of Multiscale Processes

Multiscale processes possess an important Markov property, stemming directly from the whiteness of $w(s)$. We here describe a special form of this property, closely related to our main focus in this paper, namely the finest-scale of multiscale processes. To proceed, we associate with each tree node s a set $\mathcal{F}_s$, where $\mathcal{F}_s$ contains all of the finest-scale nodes that descend from s. We also associate with each node s the random vectors ξ_s and ξ_{s^c}. The random vector ξ_s contains the elements of the set $\{x(\sigma);\ \sigma \in \mathcal{F}_s\}$, stacked into a vector, while ξ_{s^c} contains the elements of the complementary set $\{x(\sigma);\ \sigma \in \mathcal{F}_0\} \cap \{x(\sigma);\ \sigma \in \mathcal{F}_s\}^c$, stacked into a vector. It will sometimes prove convenient

[3] The mean of $x(\cdot)$ can be set to any arbitrary value, by suitably adjusting the mean of $x(0)$ and $w(\cdot)$.

to denote ξ_{s^c} by $\xi_{s\alpha_{q+1}}$; we freely use both forms. To relate ξ_s and $x(\sigma)$, we introduce the matrix $H_{s|\sigma}$, where $H_{s|\sigma}x(\sigma)$ is the linear least-squares estimate of ξ_s, given $x(\sigma)$. These conventions are illustrated in Figure 1.

The Markov property, as it relates explicitly to the finest scale, can now be stated as follows:

$$\xi_0 \;=\; \begin{pmatrix} \xi_{s\alpha_1} \\ \xi_{s\alpha_2} \\ \vdots \\ \xi_{s\alpha_{q+1}} \end{pmatrix} \;=\; \begin{pmatrix} H_{s\alpha_1|s} \\ H_{s\alpha_2|s} \\ \vdots \\ H_{s\alpha_{q+1}|s} \end{pmatrix} x(s) \;+\; \begin{pmatrix} \tilde{\xi}_{s\alpha_i|s} \\ \tilde{\xi}_{s\alpha_2|s} \\ \vdots \\ \tilde{\xi}_{s\alpha_{q+1}|s} \end{pmatrix}, \tag{2}$$

with

$$x(s),\ \tilde{\xi}_{s\alpha_1|s}, \tilde{\xi}_{s\alpha_2|s}, \ldots, \tilde{\xi}_{s\alpha_{q+1}|s} \text{ uncorrelated.} \tag{3}$$

We use this property to relate the dimension of $x(s)$ to the correlation among the vectors $\xi_{s\alpha_1}$, $\xi_{s\alpha_2}, \ldots, \xi_{s\alpha_{q+1}}$. Towards this end, (2) and (3) together imply that

$$P_{\xi_{s\alpha_i}\xi_{s\alpha_j}} \;=\; H_{s\alpha_i|s}P_{x(s)}H^T_{s\alpha_j|s} \qquad (i \neq j). \tag{4}$$

(Here and elsewhere, we adhere to the notational convention that P_x is the covariance of random vector x and P_{xy} is the cross-covariance of random vectors x and y). By elementary linear algebra [18], the rank of the cross-covariance in (4) is upper-bounded by the rank of $P_{x(s)}$, which in turn is upper-bounded by the dimension of $x(s)$. The following proposition thus follows:

Proposition 1

$$\mathit{dimension}(x(s)) \;\geq\; \max_{i \neq j} \mathit{rank}\left(P_{\xi_{s\alpha_i}\xi_{s\alpha_j}}\right).$$

If the finest-scale covariance P_{ξ_0} must match exactly some prespecified covariance, then this proposition provides a lower bound on the required multiscale state dimension. In the rather likely case that the involved cross-covariance matrices have full rank, this dimension constraint becomes quite stringent. Thus, to keep the multiscale estimation algorithm computationally efficient, we find considerable motivation to turn to reduced-order (approximate) realizations.

2.3 The Generalized Correlation Coefficient

For the purposes of developing reduced-order models, it will prove convenient to have a scalar measure of the correlation among a collection of random vectors. We thus introduce a so-called *generalized correlation coefficient*. In keeping with standard conventions, we define as follows the correlation coefficient $\rho(\eta_1, \eta_2)$ between two scalar valued random variables η_1 and η_2:

$$\rho(\eta_1, \eta_2) \equiv \begin{cases} \frac{P_{\eta_1 \eta_2}}{\sqrt{P_{\eta_1} P_{\eta_2}}} & \text{if } P_{\eta_i} > 0, \text{ for } i = 1, 2, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Then, for a pair of vector-valued random variables η_1 and η_2, we define their generalized correlation coefficient $\bar{\rho}(\eta_1, \eta_2)$ by

$$\bar{\rho}(\eta_1, \eta_2) \equiv \max_{f_1, f_2} \left\{ \rho(f_1^T \eta_1, f_2^T \eta_2) \right\}$$

where the dummy argument f_i (for $i = 1, 2$) is a column vector having the same dimension as η_i. Finally, we extend the definition of $\bar{\rho}(\cdot, \cdot)$ to a collection of random vectors $\eta_1, \eta_2, \ldots, \eta_k$, in the following way:

$$\bar{\rho}(\eta_1, \eta_2, \ldots, \eta_k) \equiv \max_{i \neq j} \bar{\rho}(\eta_i, \eta_j).$$

Each of these correlation coefficients has a conditioned version. To define them, we first let random vector z contain the conditioning information. Also, we let $\tilde{\eta}_i \equiv \eta_i - E(\eta_i | z)$, where (here and elsewhere) we adhere to the convention that $E(x|y)$ is the linear least-squares estimate of x given y. Finally, we define

$$\bar{\rho}(\eta_1, \eta_2, \ldots, \eta_k \mid z) \equiv \bar{\rho}(\tilde{\eta}_1, \tilde{\eta}_2, \ldots, \tilde{\eta}_k). \tag{6}$$

3 Formulation and Initial Investigation of Realization Problem

The realization problem of interest in this paper is to build a multiscale model, indexed on a given tree structure, to realize some pre-specified, finest-scale covariance. We begin with a random vector χ_0, having the pre-specified covariance P_{χ_0} and having an established correspondence with the finest scale of the given tree. For example, χ_0 might be a random field (written for simplicity as a vector), with the finest scale of the tree (e.g., a quadtree) being a pixel-by-pixel representation of the field. Our objective is to specify values for the model parameters $P(0)$, $A(\cdot)$ and $Q(\cdot)$, to achieve the best match possible between the actual, realized covariance P_{ξ_0} and the desired covariance P_{χ_0}. Because the desirable model properties of low dimension and high fidelity are typically in conflict, we impose a dimension constraint: for all s, the state vector $x(s)$ is constrained to have dimension no greater than λ_s:

$$\text{dimension}\,(x(s)) \;\; \geq \;\; \lambda_s. \tag{7}$$

3.1 Full-Order, Exact Realizations

When the dimension constraint is discarded, the realization problem becomes conceptually simpler and exact realizations (i.e., realizations for which $P_{\xi_0} = P_{\chi_0}$) become possible. We begin by analyzing this case.

A notable class of multiscale processes in this context is those in which each state variable $x(s)$ is a linear function of the finest-scale process:

$$x(s) \;\; = \;\; W_s \xi_s. \tag{8}$$

A state vector $x(s)$ obeying this relationship can clearly be seen to represent an aggregate (coarse) description of the finest-scale process descending from s. We refer to the matrix W_s associated with node s as the node's *internal* matrix, and we refer to multiscale processes for which (8) holds $\forall s$ as *internal* realizations. The notion of internal stochastic realizations is standard in time-series analysis [10, 16], with our use of the concept representing a natural adaptation.

Our interest in internal realizations stems from the convenient fact that the model parameters $P(0)$, $A(\cdot)$, and $Q(\cdot)$ can be specified completely in terms of the internal matrices and the finest-scale covariance. In other words, once values values for the internal matrices have been determined, values for the model parameters $P(0)$, $A(\cdot)$ and $Q(\cdot)$ follow easily. To see this fact, we begin by substituting (8) evaluated at $s = 0$ into $P(0) = E\left[x(0)x^T(0)\right]$ to yield

$$P(0) = W_0 P_{\xi_0} W_0^T. \tag{9}$$

The parameters $A(s)$ and $Q(s)$ can then be computed by noting that (1) represents the linear least-squares prediction of $x(s)$ based upon $x(s\bar{\gamma})$, plus the associated prediction error:

$$x(s) = E\left[x(s) \mid x(s\bar{\gamma})\right] + w(s) \tag{10}$$

Comparing (1) and (10), and using standard results from linear least-squares estimation, the model parameters can be seen to satisfy the following relations:

$$A(s) = P_{x(s)x(s\bar{\gamma})} P_{x(s\bar{\gamma})}^{-1} \tag{11a}$$

$$Q(s) = P_{x(s)} - P_{x(s)x(s\bar{\gamma})} P_{x(s\bar{\gamma})}^{-1} P_{x(s)x(s\bar{\gamma})}^T. \tag{11b}$$

Finally, again using (8), the covariances appearing in (11) can be expressed as simple functions of the internal matrices and the finest-scale covariance:

$$P_{x(s)x(s\bar{\gamma})} = W_s P_{\xi_s \xi_{s\bar{\gamma}}} W_{s\bar{\gamma}}^T \tag{12a}$$

$$P_{x(s)} = W_s P_{\xi_s} W_s^T. \tag{12b}$$

Clearly, the key to constructing an exact, internal realization of a given finest-scale covariance is to devise the internal matrices. At the finest-scale nodes, these matrices are implicitly defined by the association between finest-scale nodes and the components of χ_0; for example, if each scalar component of χ_0 is assigned to a distinct finest-scale node, then clearly $W_s = 1$ for each finest-scale node. At the coarse-scale nodes (i.e., all the nodes above the finest scale), the Markov property of multiscale processes becomes key. In particular, a necessary condition for (2) and (3) to hold at a coarse-scale node s in an exact, internal model is that W_s fulfill the following decorrelating role:

$$\bar{\rho}\left(\chi_{s\alpha_1}, \chi_{s\alpha_2}, \dots, \chi_{s\alpha_{q+1}} \mid W_s \chi_s\right) = 0. \tag{13}$$

In essence, the rows of W_s must contain suitable linear combinations of the random vector χ_s, such that conditioned on $W_s\xi_s$, the random vectors $\chi_{s\alpha_1}, \dots, \chi_{s\alpha_{q+1}}$ are all uncorrelated. Conversely, suppose that for a desired covariance P_{χ_0} a matrix W_s satisfying (13) is found for each coarse-scale node.[4] Suppose, further, that the resulting W_s matrices are substituted into (9), (11) and (12) to calculate values of $P(0)$, $A(\cdot)$ and $Q(\cdot)$. Then the resulting multiscale model will have the desired finest-scale covariance (i.e., $P_{\xi_0} = P_{\chi_0}$).

In summary, there is a three-stage procedure for realizing exactly any desired finest-scale covariance: (i) establish a correspondence between finest-scale nodes and components of the vector χ_0, thereby implicitly specifying W_s for each finest-scale node, (ii) find a matrix W_s satisfying (13) for every coarse-scale node, and finally (iii) substitute the resulting W_s values into (9), (11) and (12) to calculate values for $P(0)$, $A(\cdot)$ and $Q(\cdot)$. A very attractive feature of this procedure is that it decomposes the realization problem into a collection of independent sub-problems, each myopically focused on determining the information content of the state vector at a single node to fulfill the decorrelating role (13). We hasten to add, however, that the resulting state vectors will typically have an impractically high dimension, and thus this construction is mainly of interest for motivating our approach to reduced-order modeling.

3.2 Reduced-Order, Approximate Realizations

For the rest of the paper, we reinstate the constraint (7) on multiscale model state dimension. With this constraint in effect, Proposition 1 shows that exact equality between P_{ξ_0} and P_{χ_0} will in general be impossible to achieve. Therefore, we no longer look for W_s matrices that fulfill the decorrelation condition (13) exactly; instead, we look for matrices that do the best decorrelation job possible, subject to the dimension constraint.

To describe the W_s condition we use in lieu of (13), we must first introduce some notation. We define the random vectors χ_s and χ_{s^c} to have the same relation to χ_0 as ξ_s and ξ_{s^c} have to ξ. To be more precise, suppose that the i-th component of ξ_s (ξ_{s^c}) maps to the $n_s(i)$-th ($n_{s^c}(i)$-th) component of ξ_0; then, the i-th component of χ_s (χ_{s^c}) maps to the $n_s(i)$-th ($n_{s^c}(i)$-th) component

[4] The choice $W_s = I$, so that $x(s) = \xi_s$ is universally valid, though of virtually no practical value, owing to the high dimension for $x(s)$ to which it leads.

of χ_0. It will sometimes prove convenient to denote χ_{s^c} by $\chi_{s\alpha_{q+1}}$; we freely use both forms.

Now, in lieu of (13), we seek W_s matrices satisfying

$$W_s = \arg \min_{W \in \mathcal{M}_{\lambda_s}} \bar{\rho}\left(\chi_{s\alpha_1}, \chi_{s\alpha_2}, \ldots, \chi_{s\alpha_{q+1}} \mid W\chi_s\right), \tag{14}$$

where $\mathcal{M}_{\lambda_s}$ is the set of matrices having λ_s or fewer rows (and a number of columns implicitly defined by context). We refer to (14) as the *decorrelation problem*. Once values for the W_s matrices have been found, we mimic our approach to the full-order realization problem: values for the multiscale parameters $P(0)$, $A(\cdot)$ and $Q(\cdot)$ are set using analogues to (9), (11), and (12) in which P_{ξ_s} is replaced by P_{χ_s} and $P_{xi_s\xi_{s\bar{\gamma}}}$ is replaced by $P_{chi_s\chi_{s\bar{\gamma}}}$. Thus, our reduced-order modeling procedure is very similar to our three-stage, full-order modeling procedure (see Section 3.1), with the principal exception being that now condition (14) is used in lieu of condition (13).

There are several comments to make about this modeling approach. First, the approach shares with its full-order counterpart the computational benefit that we can find all the model parameters in a single sweep from coarse to fine scales, determining W_s for each node as we go along, and thereby implicitly specifying $P(0)$, $A(\cdot)$ and $Q(\cdot)$. We emphasize, though, that the condition (14) is a heuristic one. Certainly, this condition is reasonable, from a myopic, node-by-node view of the realization problem; however, the condition *does not* provide tight control over the overall match between P_{ξ_0} and P_{χ_0}. Indeed, an open research challenge is to find a way to build a reduced-order model, in which the parameters $P(0)$, $A(\cdot)$ and $Q(\cdot)$ are chosen explicitly to minimize some global measure of the discrepancy between P_{χ_0} and P_{ξ_0}. This problem appears to be very challenging. We will focus only on the more myopic problem of solving (14).

As an additional comment, models constructed with our approach will not in general be internal realizations. In other words, (8) will not hold in general. Consequently, in reduced-order models, the W_s matrices should be interpreted as merely auxiliary constructs, which aid in setting values for the parameters $P(0)$, $A(\cdot)$ and $Q(\cdot)$.

Finally, the definition of the generalized correlation coefficient makes it clear that for any given matrix W_s,

$$\bar{\rho}(\chi_{s\alpha_1}, \chi_{s\alpha_2}, \dots, \chi_{s\alpha_{q+1}} \mid W_s\chi_s) = \bar{\rho}(\chi_{s\alpha_1}, \chi_{s\alpha_2}, \dots, \chi_{s\alpha_{q+1}} \mid W_s'\chi_s)$$

where W_s' is a matrix whose rows form an orthonormal basis for the row space of W_s. Hence, defining the set $\mathcal{N}_\lambda$ to be the subset of $\mathcal{M}_\lambda$ having orthonormal rows, we see that

$$\min_{W\in\mathcal{M}_{\lambda_s}} \bar{\rho}(\chi_{s\alpha_1}, \chi_{s\alpha_2}, \dots, \chi_{s\alpha_{q+1}} \mid W\chi_s) = \min_{W\in\mathcal{N}_{\lambda_s}} \bar{\rho}(\chi_{s\alpha_1}, \chi_{s\alpha_2}, \dots, \chi_{s\alpha_{q+1}} \mid W\chi_s).$$

Thus, without loss of optimality, we can replace the constraint set $\mathcal{M}_{\lambda_s}$ in (14) with the set $\mathcal{N}_{\lambda_s}$. When convenient, we will freely make this replacement.

4 Decorrelating Sets of Random Vectors

4.1 Decorrelating a Pair of Random Vectors

We here analyze a special case of the decorrelation problems in which there are only two vectors to decorrelate. Denoting these vectors by η_1 and η_2 and stacking them as $\eta = \left(\eta_1^T \ \eta_2^T\right)^T$, our objective is to find the optimal matrix solution to the following optimization problem:

$$W^* = \arg\min_{W\in\mathcal{M}_\lambda} \bar{\rho}\,(\eta_1, \eta_2 \mid W\eta). \tag{15}$$

Playing a central role in the solution is a standard result from canonical correlation theory. For the purposes of stating this result precisely, we denote the rank of the $n_i \times n_i$ covariance matrix P_{η_i} by m_i (for $i = 1, 2$), and the rank of $P_{\eta_1\eta_2}$ by m_{12}. Also, we let I_n be an identity matrix having n rows and columns.

Theorem 1 *There exist matrices T_1 and T_2, of dimension $m_1 \times n_1$ and $m_2 \times n_2$, respectively, such that*

$$\underbrace{\begin{pmatrix} T_1 & \mathbf{0} \\ \mathbf{0} & T_2 \end{pmatrix}}_{T} \begin{pmatrix} P_{\eta_1} & P_{\eta_1\eta_2} \\ P_{\eta_1\eta_2}^T & P_{\eta_2} \end{pmatrix} \begin{pmatrix} T_1 & \mathbf{0} \\ \mathbf{0} & T_2 \end{pmatrix}^T = \begin{pmatrix} I_{m_1} & D \\ D^T & I_{m_2} \end{pmatrix},$$

and

$$\begin{pmatrix} T_1^+ & \mathbf{0} \\ \mathbf{0} & T_2^+ \end{pmatrix} \begin{pmatrix} I_{m_1} & D \\ D^T & I_{m_2} \end{pmatrix} \begin{pmatrix} T_1^+ & \mathbf{0} \\ \mathbf{0} & T_2^+ \end{pmatrix}^T = \begin{pmatrix} P_{\eta_1} & P_{\eta_1 \eta_2} \\ P_{\eta_1 \eta_2}^T & P_{\eta_2} \end{pmatrix}.$$

The matrix D has dimension $m_1 \times m_2$ and is given by $D = diag\left(\hat{D}, \mathbf{0}\right)$, where $\hat{D} = diag(d_1, d_2, ..., d_{m_{12}})$, $1 \geq d_1 \geq d_2 \geq ... \geq d_{m_{12}} > 0$: for a given $(P_{\eta_1}, P_{\eta_2}, P_{\eta_1 \eta_2})$, the matrix $\hat{D}$ is unique. Finally, T_i^+ is the Moore-Penrose pseudoinverse of T_i, and is given by $T_i^+ = P_{\eta_i} T_i^T$, $(i = 1, 2)$.

We refer to the triple of matrices (T_1, T_2, D) as the *canonical correlation matrices* associated with (η_1, η_2). For convenience, we introduce truncated versions of T_i (for $i = 1, 2$), denoted by $T_{i,k}$ and defined to contain the first k rows of T_i; as a special case, we define $\hat{T}_i$ to contain the first m_{12} rows of T_i. Results very similar to Theorem 1 can be found in several places, including [3, 14, 15] and [6]. A proof of the theorem, as exactly stated here, can be found in [9]. As these proofs reveal, the calculation of the canonical correlation matrices can be carried out in a numerically sound fashion using the singular value decomposition; this calculation requires $\mathcal{O}(N^3)$ floating point operations, where $N = \max(n_1, n_2)$.

Theorem 1 can be used to perform a change of basis on the vectors η_1 and η_2, to simplify maximally the correlation between them, and thus to simplify analysis of the decorrelation problem. We define the random vectors μ, μ_1 and μ_2 via

$$\mu \equiv \begin{pmatrix} \mu_1^T & \mu_2^T \end{pmatrix}^T, \quad \mu_i = T_i \, \eta_i, \qquad (i = 1, 2),$$

where thanks to Theorem 1, μ_1 and μ_2 have covariance

$$P_\mu = \begin{pmatrix} P_{\mu_1} & P_{\mu_1 \mu_2} \\ P_{\mu_1 \mu_2}^T & P_{\mu_2} \end{pmatrix} = \begin{pmatrix} I_{m_1} & D \\ D^T & I_{m_2} \end{pmatrix},$$

and the transformation from (η_1, η_2) to (μ_1, μ_2) is invertible in a mean-square sense,

$$E[(\eta_i - T_i^+ \mu_i)(\eta_i - T_i^+ \mu_i)^T] = \mathbf{0}, \qquad (i = 1, 2).$$

The following lemma now provides the key simplification.

Lemma 1

$$\bar{\rho}(\eta_1,\eta_2 \mid WT\eta) = \bar{\rho}(\mu_1,\mu_2 \mid W\mu) \tag{16a}$$

$$\bar{\rho}(\eta_1,\eta_2 \mid W\eta) = \bar{\rho}\,(\mu_1,\mu_2 \mid WT^{+}\mu) \tag{16b}$$

$$\min_{W\in\mathcal{M}_\lambda} \bar{\rho}(\eta_1,\eta_2 \mid W\eta) = \min_{V\in\mathcal{N}_\lambda} \bar{\rho}(\mu_1,\mu_2 \mid V\mu) \tag{16c}$$

As a special case of (16a) and (16b), we note that $\bar{\rho}(\eta_1,\eta_2) = \bar{\rho}(\mu_1,\mu_2)$. The lemma is a direct consequence of the definition of the generalized correlation coefficient, together with Theorem 1; we omit the details of the proof.

Equipped with the foregoing theorem and lemma, we can now solve (15).

Proposition 2 *For* $0 \leq \lambda < m_{12}$ *and for* $i = 1,2$,

$$\min_{W\in\mathcal{M}_\lambda} \bar{\rho}(\eta_1,\eta_2 \mid W\eta) = \min_{W_1\in\mathcal{M}_\lambda} \bar{\rho}(\eta_1,\eta_2 \mid W_1\eta_1) \;=\; \bar{\rho}(\eta_1,\eta_2 \mid T_{1,\lambda}\eta_1) \;=\; d_{\lambda+1}. \tag{17a}$$

For $\lambda \geq m_{12}$,

$$\min_{W\in\mathcal{M}_\lambda} \bar{\rho}(\eta_1,\eta_2 \mid W\eta) = \min_{W_1\in\mathcal{M}_\lambda} \bar{\rho}(\eta_1,\eta_2 \mid W_1\eta_1) \;=\; \bar{\rho}(\eta_1,\eta_2 \mid \hat{T}_1\eta_1) \;=\; 0. \tag{17b}$$

Proof: In Appendix 1, we demonstrate that for $\lambda < m_{12}$,

$$\min_{W\in\mathcal{N}_\lambda} \bar{\rho}(\mu_1,\mu_2 \mid W\mu) = \min_{W_1\in\mathcal{N}_\lambda} \bar{\rho}(\mu_1,\mu_2 \mid W_1\mu_1) \;=\; \bar{\rho}(\mu_1,\mu_2 \mid \begin{pmatrix} I_\lambda & \mathbf{0} \end{pmatrix} \mu_1) \;=\; d_{\lambda+1}, \tag{18a}$$

while for $\lambda \geq m_{12}$,

$$\min_{W\in\mathcal{N}_\lambda} \bar{\rho}(\mu_1,\mu_2 \mid W\mu) = \min_{W_1\in\mathcal{N}_\lambda} \bar{\rho}(\mu_1,\mu_2 \mid W_1\mu_1) \;=\; \bar{\rho}(\mu_1,\mu_2 \mid \begin{pmatrix} I_{m_{12}} & \mathbf{0} \end{pmatrix} \mu_1) \;=\; 0. \tag{18b}$$

Once these facts are established, the results (17a) and (17b) then follow. In particular, with regard to (17a), we have the following sequence of identities:

$$\begin{aligned} \min_{W\in\mathcal{M}_\lambda} \bar{\rho}(\eta_1,\eta_2 \mid W\eta) \;&=\; \min_{W\in\mathcal{N}_\lambda} \bar{\rho}(\mu_1,\mu_2 \mid W\mu) \;=\; \bar{\rho}(\mu_1,\mu_2 \mid \begin{pmatrix} I_\lambda & \mathbf{0} \end{pmatrix} \mu_1) \\ &=\; \bar{\rho}(\eta_1,\eta_2 \mid \begin{pmatrix} I_\lambda & \mathbf{0} \end{pmatrix} T_1\eta_1) \;=\; d_{\lambda+1}. \end{aligned} \tag{19}$$

The first equality follows from (16c), the second from (18a), the third from (16a) and the fourth from (18a). The result (17b) can be proved from (18b) in a very similar fashion: the details are omitted. **QED**.

There are two important points to note about this proposition. The first is that solving (15) is essentially a problem of calculating the canonical correlation matrices associated with (η_1, η_2); indeed, (15) can be solved simultaneously for all values of λ by calculating just once these canonical correlation matrices. The second point is that there is no harm in having the decorrelating information $W\eta$ be a linear function of either η_1 or η_2 alone.

4.2 Decorrelating Multiple Random Vectors

We now turn to the general decorrelation problem (14), for which we develop a suboptimal solution. This solution has an intuitively appealing structure motivated by the solution to the simpler problem (15). We emphasize that to the best of our knowledge, the task of characterizing the *optimal* solution to (14) is an unsolved problem.

Our approach is to decompose the decorrelation problem into a collection of q sub-problems. In the i-th sub-problem, we focus on decorrelating $\chi_{s\alpha_i}$ from $\chi_{s\alpha_j}$ for all $j \neq i$; specifically, we exploit Proposition 2 to solve

$$W_{i,k_i} \;=\; \arg \min_{W \in \mathcal{M}_{k_i}} \bar{\rho}\left(\chi_{s\alpha_i}, \chi_{(s\alpha_i)^c} \mid W\chi_{s\alpha_i}\right), \tag{20}$$

where for now we treat $k_1, \ldots , k_q$ as free parameters. By choosing W_{i,k_i} as in (20), we effectively decorrelate $\chi_{s\alpha_i}$ from $\chi_{s\alpha_j}$ (for all $j \neq i$) all at once; in particular, it is clear that

$$\bar{\rho}\left(\chi_{s\alpha_i}, \chi_{s\alpha_j} \mid W_{i,k_i}\chi_{s\alpha_i}\right) \;\leq\; \bar{\rho}\left(\chi_{s\alpha_i}, \chi_{(s\alpha_i)^c} \mid W_{i,k_i}\chi_{s\alpha_i}\right), \quad j \neq i, \tag{21}$$

and so, if the right side of (21) is small, then the left side will also be for all $j \neq i$.

To see how we combine $W_{1,k_1}, \ldots , W_{q,k_q}$ to solve (14) approximately, let us consider the quantity

$$\bar{\rho}\left(\chi_{s\alpha_1}, \ldots , \chi_{s\alpha_q} \mid W_{1,k_1}\chi_{s\alpha_1}, \ldots , W_{q,k_q}\chi_{s\alpha_q}\right), \tag{22}$$

which we can express more succinctly as $\bar{\rho}(\chi_{s\alpha_1}, \ldots, \chi_{s\alpha_q} | W_s(k_1, \ldots, k_q)\chi_s)$ by defining the block-diagonal matrix $W_s(k_1, \ldots, k_q) \equiv \mathrm{diag}(W_{1,k_1}, \ldots, W_{q,k_q})$. Since the i-th block component of this matrix has been specially designed to decorrelate $\chi_{s\alpha_i}$ from $\chi_{s\alpha_j}$, $j \neq i$, we intuitively expect that all the block components will work together to make (22) small. Furthermore, if

$$\sum_{i=1}^{q} k_i \;\leq\; \lambda_s, \tag{23}$$

then $W_s(k_1, \ldots, k_q) \in \mathcal{M}_{\lambda_s}$, implying that $W_s(k_1, \ldots, k_q)$ is in the feasible set of the optimization problem (14), and can indeed be used as an approximate solution to (14).

To characterize precisely the behavior of $W_s(k_1, \ldots, k_q)$, we first must establish a result describing the non-increasing nature of the generalized correlation coefficient as the amount of conditioning information increases:

Proposition 3

$$\bar{\rho}(\eta_1, \eta_2 \mid W_i\eta_i) \;\leq\; \bar{\rho}(\eta_1, \eta_2), \quad i = 1, 2. \tag{24}$$

Proof: In Appendix B, we demonstrate that

$$\bar{\rho}\,(\mu_1, \mu_2 \mid W_i\mu_i) \;\leq\; \bar{\rho}\,(\mu_1, \mu_2)\,. \tag{25}$$

Once this fact is established, the result (24) follows. In particular, we have the following sequence of relations:

$$\bar{\rho}\,(\eta_1, \eta_2 \mid W_i\eta_i) \;=\; \bar{\rho}\,(\mu_1, \mu_2 \mid W_iT_i^{+}\mu_i) \;\leq\; \bar{\rho}\,(\mu_1, \mu_2) \;=\; \bar{\rho}\,(\eta_1, \eta_2)\,.$$

The first relation follows from (16b), the second from (25) and the third from (16a). **QED**.

We emphasize that if the conditioning information is not a function of either η_1 or η_2 alone, then the function $\bar{\rho}(\cdot, \cdot \mid \cdot)$ may become an increasing one. For instance, if

$$\begin{pmatrix} P_{\eta_1} & P_{\eta_1\eta_2} \\ P_{\eta_1\eta_2} & P_{\eta_2} \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

then $\bar{\rho}(\eta_1, \eta_2) = 0.5$, but $\bar{\rho}(\eta_1, \eta_2 \mid \eta_1 + \eta_2) \;=\; 1$.

We can, however, slightly strengthen Proposition 3 by relaxing our restriction that *all* of the conditioning information be a linear function of either η_1 or η_2; in lieu of this restriction, we restrict *each individual scalar component* of this conditioning information to be a function of either η_1 or η_2. We state this result as a corollary:

Corollary 1

$$\bar{\rho}(\eta_1, \eta_2 \mid W_1\eta_{i_1}, W_2\eta_{i_2}) \quad \leq \quad \bar{\rho}(\eta_1, \eta_2 \mid W_1\eta_{i_1}), \qquad (i_1, i_2) \in \{\{1,2\} \times \{1,2\}\}$$

Proof:

$$\begin{aligned} \bar{\rho}(\eta_1, \eta_2 \mid W_1\eta_{i_1}, W_2\eta_{i_2}) &= \bar{\rho}(\eta_1 - E(\eta_1|W_1\eta_{i_1}), \eta_2 - E(\eta_1|W_1\eta_{i_1}) \mid W_2(\eta_{i_2} - E(\eta_{i_2}|W_1\eta_{i_1}))) \\ &\leq \bar{\rho}(\eta_1 - E(\eta_1|W_1\eta_{i_1}), \eta_2 - E(\eta_1|W_1\eta_{i_1})) \\ &= \bar{\rho}(\eta_1, \eta_2 \mid W_1\eta_{i_1}) \end{aligned}$$

The first and third lines here represent direct applications of (6), while the second line represents application of Proposition 3. **QED**.

Using this corollary, we now return to consideration of the behavior of $W_s(k_1, \ldots, k_q)$.

Proposition 4

$$\bar{\rho}(\chi_{s\alpha_1}, \ldots, \chi_{s\alpha_{q+1}} \mid W_s(k_1, \ldots, k_q)\chi_s) \leq \max_{i=1,\ldots,q} \bar{\rho}(\chi_{s\alpha_i}, \chi_{(s\alpha_i)^c} \mid W_s(k_1, \ldots, k_q)\chi_s) \tag{26a}$$

$$\leq \max_{i=1,\ldots,q} \bar{\rho}(\chi_{s\alpha_i}, \chi_{(s\alpha_i)^c} \mid W_{s,k_i}\chi_{s\alpha_i}). \tag{26b}$$

Proof: The first inequality in (26) is a direct consequence of the definition of the generalized correlation coefficient. The second is then a direct consequence of the corollary to Proposition 3. **QED**.

The important point to note about this proposition is that $W_s(k_1, \ldots, k_q)$ leads to a value for the objective function in (14) that is no greater than the maximum of the values obtained in the q sub-problems (20). In other words, by concatenating together the solutions to the q sub-problems (20) into the block-diagonal matrix $W_s(k_1, \ldots, k_q)$, we obtain an approximate solution to (14) having a value upper-bounded by the maximum value of these q solutions to (20). This observation suggests a way to select values for the parameters $k_1, \ldots, k_q$. In particular, subject to the constraint (23), we should choose these parameters to fulfill the following minimax condition:

$$(k_1^*, \ldots, k_q^*) = \arg \min_{k_1, \ldots, k_q} \left\{ \max_{i=1,\ldots,q} \bar{\rho}(\chi_{s\alpha_i}, \chi_{(s\alpha_i)^c} \mid W_{i,k_i} \chi_{s\alpha_i}) \right\}. \tag{27}$$

By choosing the k_i parameters in this fashion, we minimize the right side of (26b), which upper-bounds the left side of (26a). The matrix $W_s(k_1^*, \ldots, k_q^*)$ then can serve as a suboptimal solution to (14).

To describe the solution to (27), we denote by $(\hat{T}_{s\alpha_i}, \hat{T}_{(s\alpha_i)^c}, \hat{D}_{s,i})$ the canonical correlation matrices associated with $(\chi_{s\alpha_i}, \chi_{(s\alpha_i)^c})$, where the diagonal elements of $\hat{D}_{s,i}$ are denoted by $d_1^{s,i}, d_1^{s,i}, \ldots$. For simplicity of exposition only, we assume that k_i is strictly less than the rank of the cross-covariance $P_{\chi_{s\alpha_i}\chi_{(s\alpha_i)^c}}$, for $i = 1, \ldots q$. Then, thanks to Proposition 2, it follows that

$$\bar{\rho}(\chi_{s\alpha_i}, \chi_{(s\alpha_i)^c} \mid W_{i,k_i} \chi_{s\alpha_i}) = d_{k_i+1}^{s,i}.$$

Hence, the minimax definition (27) is equivalent to the following, where we again impose the constraint (23):

$$(k_1^*, \ldots, k_q^*) = \arg \min_{k_1, \ldots, k_q} \left\{ \max_{i=1,\ldots,q} d_{k_i+1}^{s,i} \right\}$$

This discrete optimization problem can easily be solved, once the canonical correlation quantities $\left(\hat{T}_{s\alpha_i}, \hat{D}_{s,i}\right)$ associated with $\left(\chi_{s\alpha_i}, \chi_{(s\alpha_i)^c}\right)$ have been calculated, for $i = 1, 2, \ldots, q$ [9].

4.3 Calculating the Canonical Correlation Matrices

For problems of practical interest to us, the dimension of χ_s and χ_{s^c} can be on the order of a thousand (or greater), thus prohibiting the exact calculation of the associated canonical correlation matrices $(\hat{T}_s, \hat{D}_s)$. However, if the correlation between χ_s and χ_{s^c} has a certain special structure, then we can achieve a substantial reduction in the complexity of the computation. In essence, we assume that the correlated component of χ_s and χ_{s^c} lives in some low-dimensional subspace that is easily defined: we then do all of our computations with low-dimensional random vectors that live in this subspace, and thereby achieve our complexity reduction.

To be more precise, we introduce the random vectors μ_s and μ_{s^c} as

$$\mu_s \;=\; \Theta_s \chi_s \quad \text{and} \quad \mu_{s^c} \;=\; \Theta_{s^c} \chi_{s^c}. \tag{28}$$

Here, Θ_s and Θ_{s^c} are matrices having having full row rank and such that

$$E\left[(\chi_s - E(\chi_s|\mu))\,(\chi_{s^c} - E(\chi_{s^c}|\mu))^T\right] \;=\; \mathbf{0}, \tag{29}$$

for both $\mu = \mu_s$ *and* $\mu = \mu_{s^c}$. It is not difficult to see that if $(\hat{T}_1^\mu, \hat{T}_2^\mu, \hat{D}^\mu)$ are the canonical correlation matrices for (μ_s, μ_{s^c}), then

$$\hat{T}_s \;=\; \hat{T}_s^\mu \Theta_s, \quad \text{and} \quad \hat{D}_s \;=\; \hat{D}_s^\mu. \tag{30}$$

This result leads to considerable computational savings when χ_0 is a wide-sense Markov random process or field. Even if χ_0 represents a WSMRF as large as 256 × 256, then the canonical correlation matrices associated with (χ_s, χ_{s^c}) can be computed in a manageable fashion to machine precision. Moreover, for non-Markov processes and fields, a slight generalization of this approach serves effectively as a method for obtaining good approximate results.

We illustrate the approach by considering a 2-D example. In the example, we let χ_0 represent the values of a first-order, scalar-valued WSMRF over a discrete lattice having dimensions 256× 256. We focus on a particular node s for which χ_s and χ_{s^c} contain the values of the field at the subsets of points displayed in Figure 2a. Specifically, χ_s contains the values of the field at the 64 grid points marked with circles, both filled and not filled, in the white region, while χ_{s^c} contains the values at the all other grid points; subsets of these other grid points are marked with squares, both filled and not filled.

Thanks to the Markov property, we can devise by inspection matrices Θ_s and Θ_{s^c} to fulfill (29). In particular, we can let Θ_s and Θ_{s^c} be selection matrices chosen such that μ_s and μ_{s^c} contain the values of χ_s and χ_{s^c} at their respective boundary points, where these boundary points are marked with filled-in circles and squares, respectively. To see the computational savings that can result by

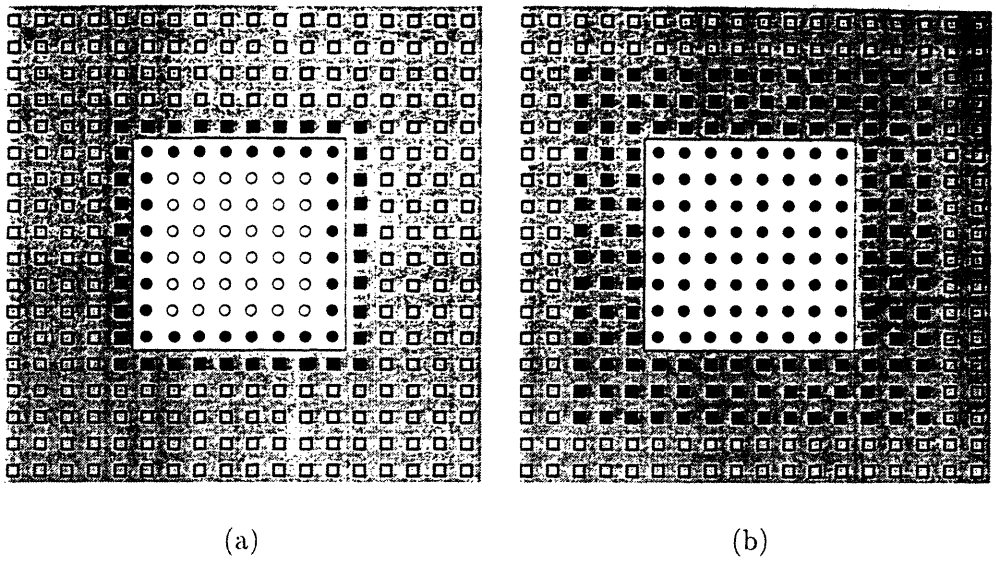


Figure 2: Illustration of our approach to finding the canonical correlation matrices associated with (χ_s, χ_{s^c}) for (a) a first-order WSMRF, and (b) a non-Markov random field.

using (30) to calculate $\hat{T}_s$ and $\hat{D}_s$. we note that the dimension of μ_{s^c} is roughly 5×10^{-4} times the dimension of χ_{s^c}, this approach reduces the computational cost of determining $(\hat{T}_s, \hat{D}_s)$ by roughly a factor of 6×10^9. From this example, the structure of our approach should be clear, for any case in which we are modeling a WSMRF.

For non-Markov random fields. there is no guarantee that the correlated component of (χ_s, χ_{s^c}) can be captured by boundary information over a region as thin as the one used in our foregoing example. To compensate for this fact, we modify our approach slightly for the non-Markov case. Our modified strategy is to make the boundary region as thick as possible for each of χ_s and χ_{s^c}, subject to the constraint that the resulting vectors μ_s and μ_{s^c} have dimension no greater than some prescribed limit. Using the same graphical conventions as in Figure 2a, this idea is illustrated in Figure 2b. where 132 is the limiting dimension of both μ_s and μ_{s^c}. Once μ_s and μ_{s^c} have been defined, we proceed exactly as in the Markov case.

5 Numerical Examples

In this section, we present two numerical examples that suggest the promise of our modeling approach. In all cases, the models we build are indexed on quadtrees. Also, for the purposes of calculating the canonical correlation matrices, we set $\Theta_{rows} = 260$.

5.1 Reduced-order Representations of Isotropic Random Fields

For our first example, we consider a scalar, wide-sense stationary, zero-mean, isotropic (but non-Markov) random field $y(m,n)$ that is of interest in the geological sciences [17]. The correlation function for this field can be expressed analytically as follows:

$$R_{yy}(i,j) \;=\; E\left[y(m+i,n+j)y(m,n)\right] \;=\; R_{yy}(r) \;=\; \begin{cases} 1 - 3/2(r/\ell) + 1/2(r/\ell)^3 & 0 \leq r \leq \ell, \\ 0 & r > \ell, \end{cases} \tag{31}$$

where $r = \sqrt{i^2 + j^2}$, and ℓ is the characteristic length of the function. A plot of this function for $\ell = 80$ is represented by the solid curve in Figure 4; we see from this plot that there is significant long-range correlation, at least relative to the total grid size we will be using.

We build multiscale models to realize the correlation function (31) on a 128 × 128 grid. We build four models, each involving a different constraint on the state dimension; we constrain the state dimension to be no greater than the respective values 64, 32, 16 and 8.

In Figure 3a, we display as a contour plot the exact correlation function (31). Then, in Figures 3b, c and d, we display as contour plots the correlation function associated with our multiscale models of order 32, 16, and 8, respectively. Because our multiscale models have reduced order, they lead to correlation structures that are only approximately stationary, and thus we must define carefully what is being plotted in Figures 3b, c and d. Towards this end, we let ξ_0^λ denote the random vector comprising the finest-scale of the particular multiscale process in which the state vectors are constrained to have dimension no greater than λ; we thus have ξ_0^8, $\xi_0^1 6$, and ξ_0^{32}. We denote the (i,j)-th component of ξ_0^λ by $\xi_0^\lambda(i,j)$ (for $i,j = 0, 1, \ldots, 127$). In terms of these conventions, the plots in Figures 3b, c, and d display contours of the function $R_\lambda(\cdot,\cdot)$, for $\lambda = 32$, 16 and

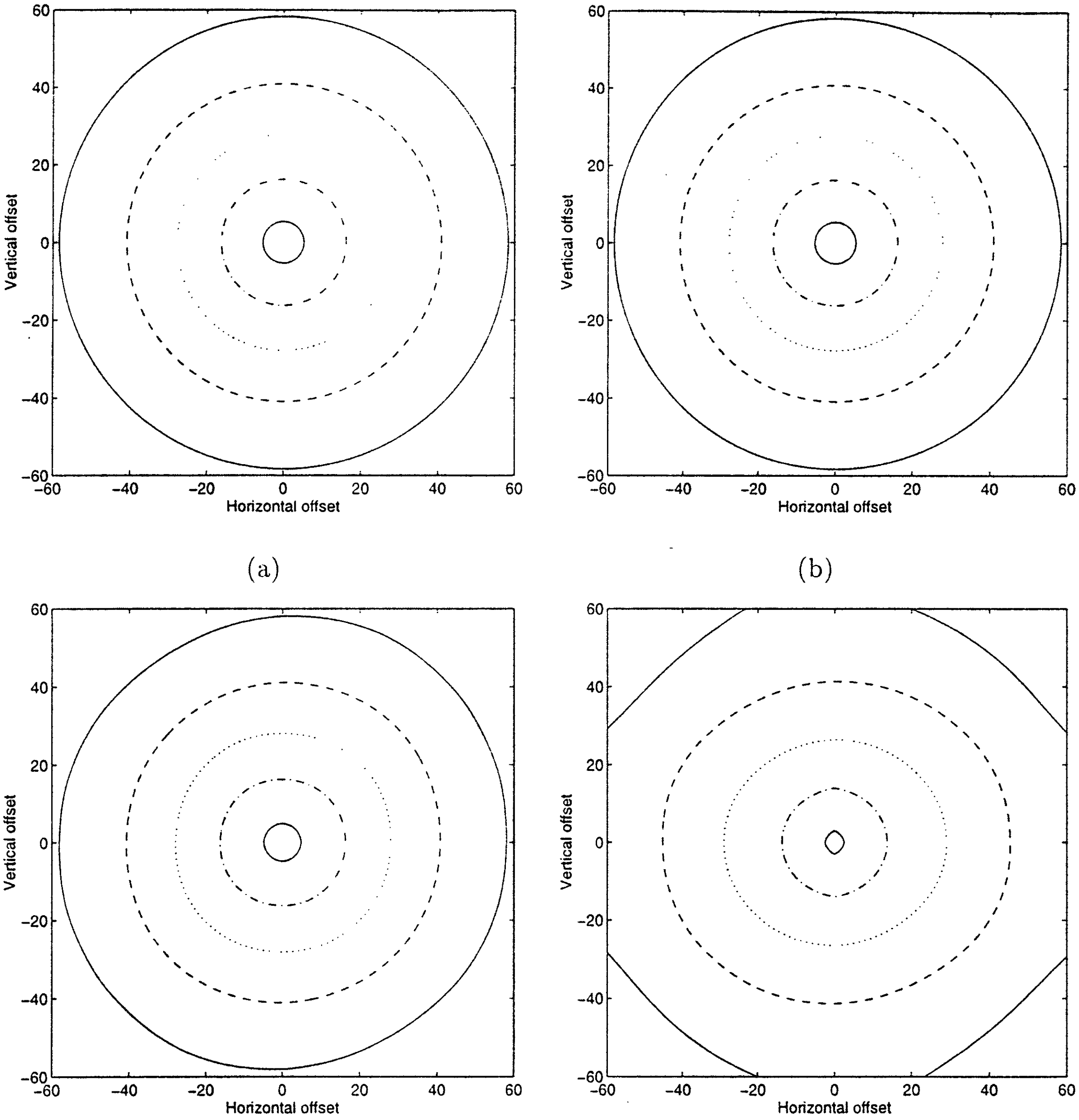
8 respectively, where

$$R_\lambda(m,n) \equiv \frac{1}{(128-m)(128-n)} \sum_{i=0}^{127-m} \sum_{j=0}^{127-n} E\left[\xi_0^\lambda(i+m,j+n)\,\xi_0^\lambda(i,j)\right]. \qquad (32)$$

We do not include a contour plot for our model of order 64, because for orders greater than just 16, our multiscale models capture virtually all of the significant correlation structure. This fact is reinforced in Figures 4a and b, where we display horizontal and vertical slices of these contour plots.

Let us consider the use of these multiscale models to carry out linear least-squares estimation. In Figure 5a, we display the original signal that we will be attempting to estimate. This signal consists of 128×128 pixels and has a Gaussian distribution. It is drawn from the *exact* distribution implied by (31) with $\ell = 80$. This field generation is effected by embedding the 128×128 grid into a larger 256×256 toroidal lattice, and extending the definition of $R_{yy}(\cdot,\cdot)$ to have periodic boundary conditions; for $\ell = 80$, this approach leads to a valid (i.e., positive definite) correlation function.

We consider two estimation problems related to the signal in Figure 5a. For the first, we corrupt the signal with spatially stationary white noise having covariance one, thus leading to an SNR of 0dB (since the signal also has a variance of one, as indicated by (31)). In Figure 5b, we display an estimate based on our multiscale model of order 64. The sample MSE here is 0.0498. While there is no computationally feasible way to determine the mean-square error of an optimal estimator for this problem, we can obtain a fairly tight lower bound for the optimal MSE. In particular, let us consider the problem of estimating the value of the 256×256 signal, from which our 128×128 signal has been extracted. Since this 256×256 signal is stationary and is indexed on a toroidal lattice, exact calculations are possible. In particular, for estimating this signal in 0dB white noise, the optimal, FFT-based estimator has an MSE of 0.0458, which must lower-bound the MSE of an optimal estimator in our original estimation problem. By comparison, then, our measured MSE of 0.0498 is quite satisfactory. Although not shown in the Figure, the same level of performance is also achieved by our lower-order multiscale models; specifically, our models of order 32, 16 and 8



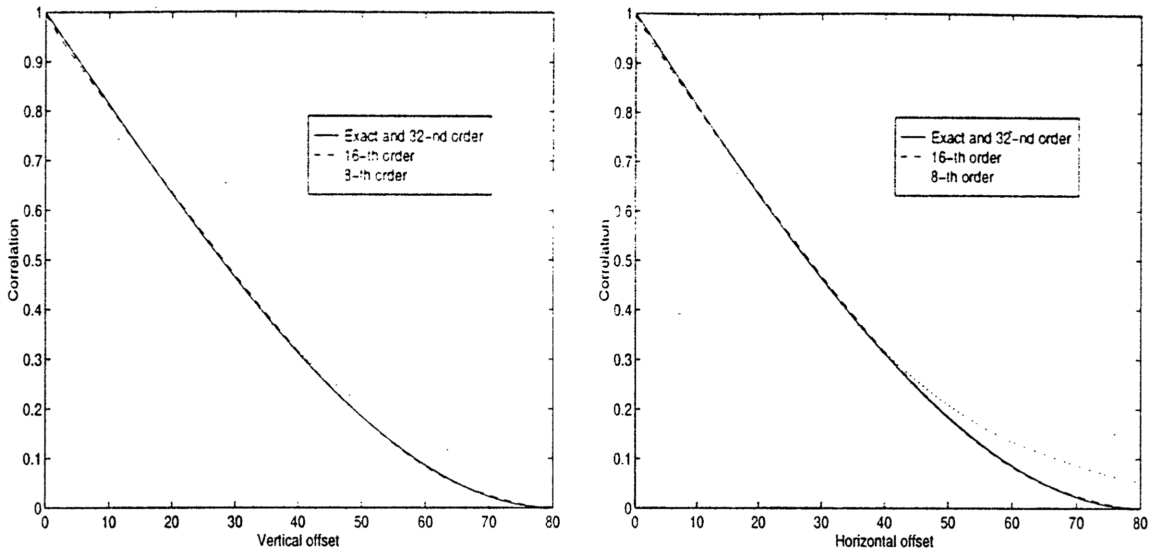
(a)

(b)

(c)

(d)

Figure 3: These four figures display contour plots associated with $R_{yy}(\cdot,\cdot)$, defined in (31), with the contour levels at 0.9, 0.7, 0.5, 0.3 and 0.1. (a) The exact, desired correlation function. (b), (c), and (d) The correlation function associated with multiscale models of order 32, 16 and 8, respectively. These three have been determined by Monte-Carlo simulation, using enough trials so that every estimated correlation value is within 0.005 of its correct value.



(a) (b)

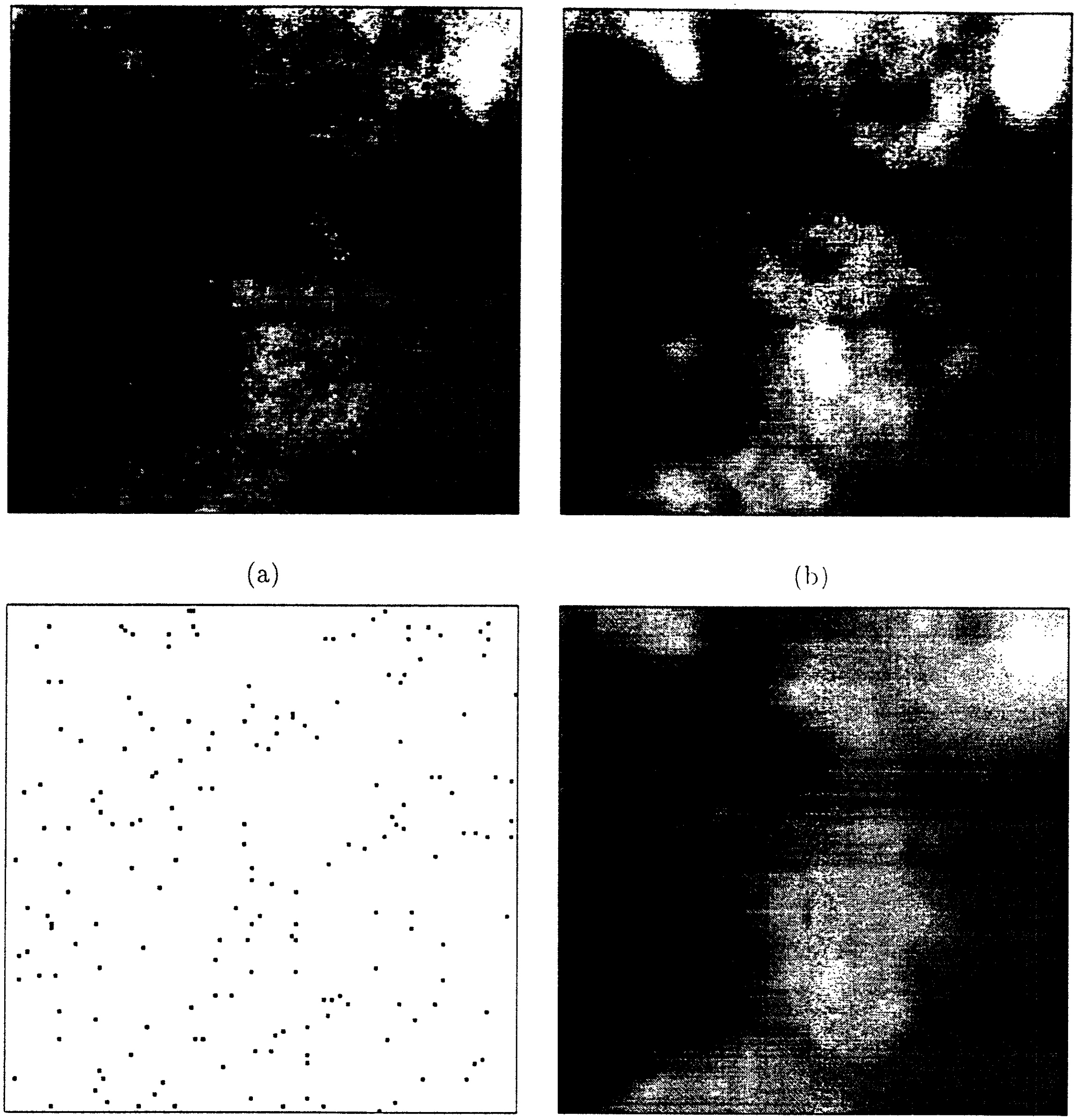
Figure 4: Comparison of (a) vertical and (b) horizontal slices of the correlation contour plots in the previous figure. Again, these plots are based on Monte-Carlo simulation. where each point is within 0.005 of its correct value with 95 percent confidence.

achieve sample MSEs of 0.0501, 0.0533 and 0.0544, respectively, which are all close to the optimal.

The second estimation problem we consider is one for which the FFT is of little practical use. In particular, we consider the problem of estimating the signal displayed in Figure 5a, based on noiseless measurements at the extremely sparse set of points displayed in Figure 5c. These points provide only 1.11% coverage of the image region. Their irregular distribution is the key reason that FFT techniques are not useful. On the other hand, in Figure 5d, we display the estimate that results from use of our multiscale model of order 64. In light of the sparsity of our measurement coverage. this estimate has impressively captured the coarse qualitative features of the true signal; in fact, the sample MSE of this estimate is only 0.1147.

5.2 Reduced-order Representations of Warped-version of Isotropic Correlation Function

For our second example, we build multiscale representations for a stationary random field having a correlation function that is a warped version of the isotropic correlation function $R_{yy}(k,l)$ in (31.



(a)

(b)

(c)

(d)

Figure 5: These four figures relate to linear least-squares estimation of a signal having the isotropic correlation function in (31). (a) The original signal, with Gaussian deviates, drawn from the exact distribution using FFT-based techniques. (b) Estimate of the sample path in (a), based on noisy, densely distributed measurements of the signal, with 0dB SNR; a 64-th order multiscale model is used to obtain this estimate. (c) Locations of observed pixels, for a second estimation experiment; these observed pixels provide only 1.11 % coverage of the image. (d) Estimate of the sample path in (a), based on noiseless observations of the observed pixels (displayed in (c)).

Our warped version, which we denote by $R'_{yy}(k,l)$ is defined as follows:

$$R'_{yy}(i,j) = R_{yy}(i',j'), \tag{33a}$$

$$\begin{pmatrix} i' \\ j' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} i \\ j \end{pmatrix}, \quad \theta = \frac{\pi}{4} - \frac{\pi}{13}. \tag{33b}$$

The characteristic length ℓ of $R_{yy}(i,j)$ (see (31)) is again set to $\ell = 80$. A contour plot of $R'_{yy}(i,j)$ is displayed in Figure 6a, while slices of this correlation function along the directions of strongest and weakest correlation are displayed in Figures 7a and b, respectively.

We consider the problem of building a multiscale model, indexed on a quadtree, to realize the correlation function (33b) on a 128×128 grid. We constrain the multiscale model dimension to the respective values of 64, 32, 16 and 8.

In Figures 6b, c and d, we display as contour plots the correlation function associated with our multiscale models of order 32, 16, and 8, respectively. Just as in our previous example, we must carefully define the precise meaning of these contours. As in the previous example, we let ξ_0^λ denote the random vector comprising the finest-scale of the particular multiscale process in which the state vectors are constrained to have dimension no greater than λ; we thus have ξ_0^8, $\xi_0^1 6$, and ξ_0^{32}. We denote the (i,j)-th component of ξ_0^λ by $\xi_0^\lambda(i,j)$ (for $i,j = 0,1,\ldots,127$). In terms of these conventions, the plots in Figures 6b, c, and d display contours of the function $R_\lambda(\cdot,\cdot)$, for $\lambda = 32$, 16 and 8 respectively, where $R_\lambda(\cdot,\cdot)$ is defined in (32). We do not include a contour plot for our model of order 64, because at this order, the contour plot is indistinguishable from the ideal, desired correlation in (a). To allow for more direct comparison of these contours, we overlay slices of them in Figures 7a and b; more specifically, Figure 7a represents a slice of the contour plots, along the direction of strongest correlation, while part b represents a slice of the contour plots along the direction of weakest correlation.

In Figure 8, we display sample paths of this random field using Gaussian deviates, generated with our models of order 64, 32, 16 and 8. We see that unless a relatively high order model is used, the sample paths exhibit visually distracting blocky artifacts at the quadrantal boundaries.

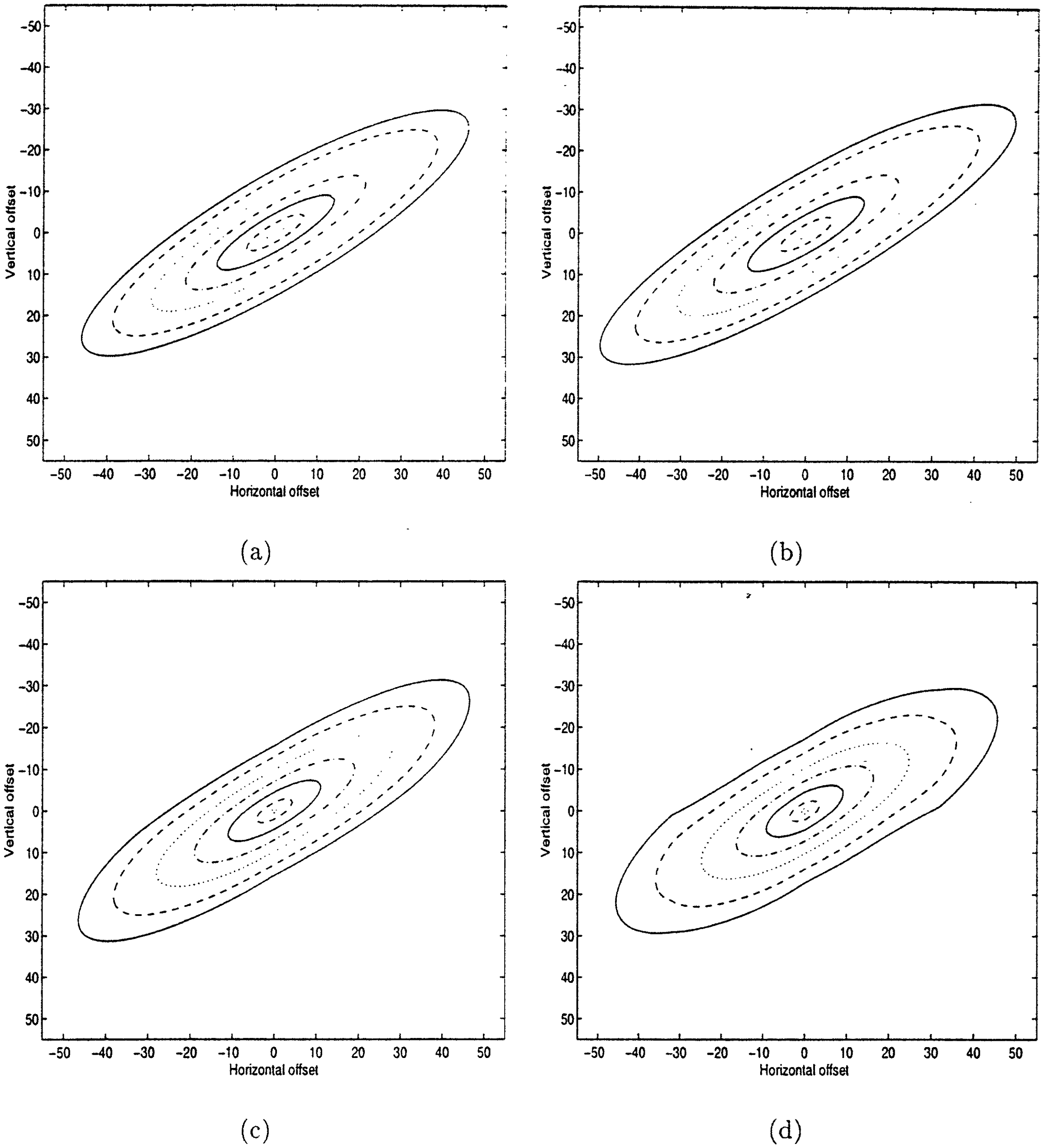
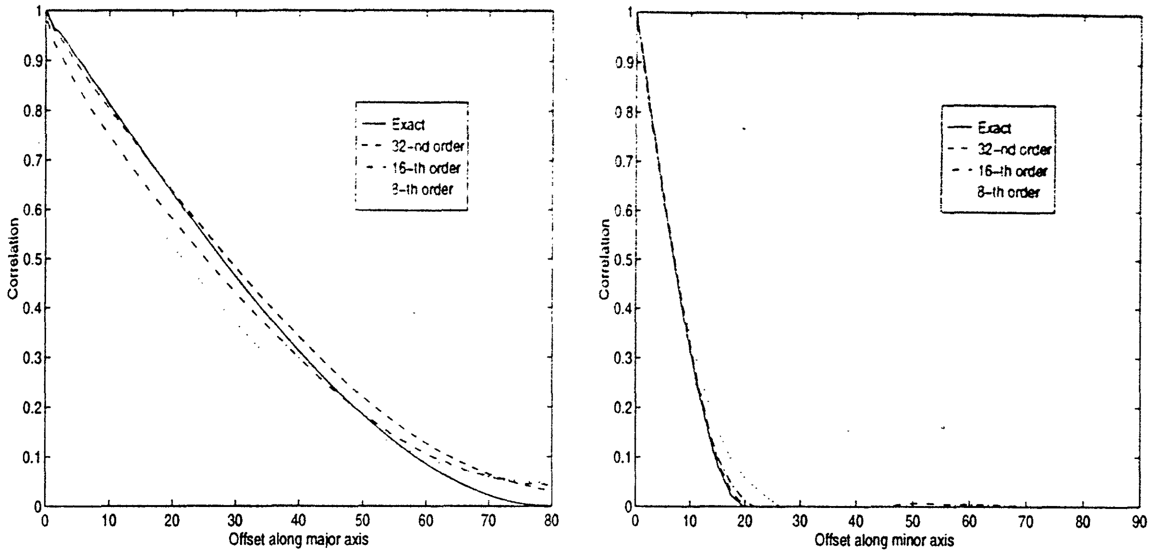


Figure 6: These four figures display contour plots associated with $R'_{yy}(\cdot,\cdot)$, defined in (33b), with the contour levels at 0.95, 0.85, 0.75, 0.6, 0.45, 0.3 and 0.15. (a) The exact, desired correlation function. (b), (c), and (d) The correlation function associated with multiscale models of order 32, 16 and 8, respectively. These three have been determined by Monte-Carlo simulation, using enough trials so that every estimated correlation value is within 0.005 of its correct value.



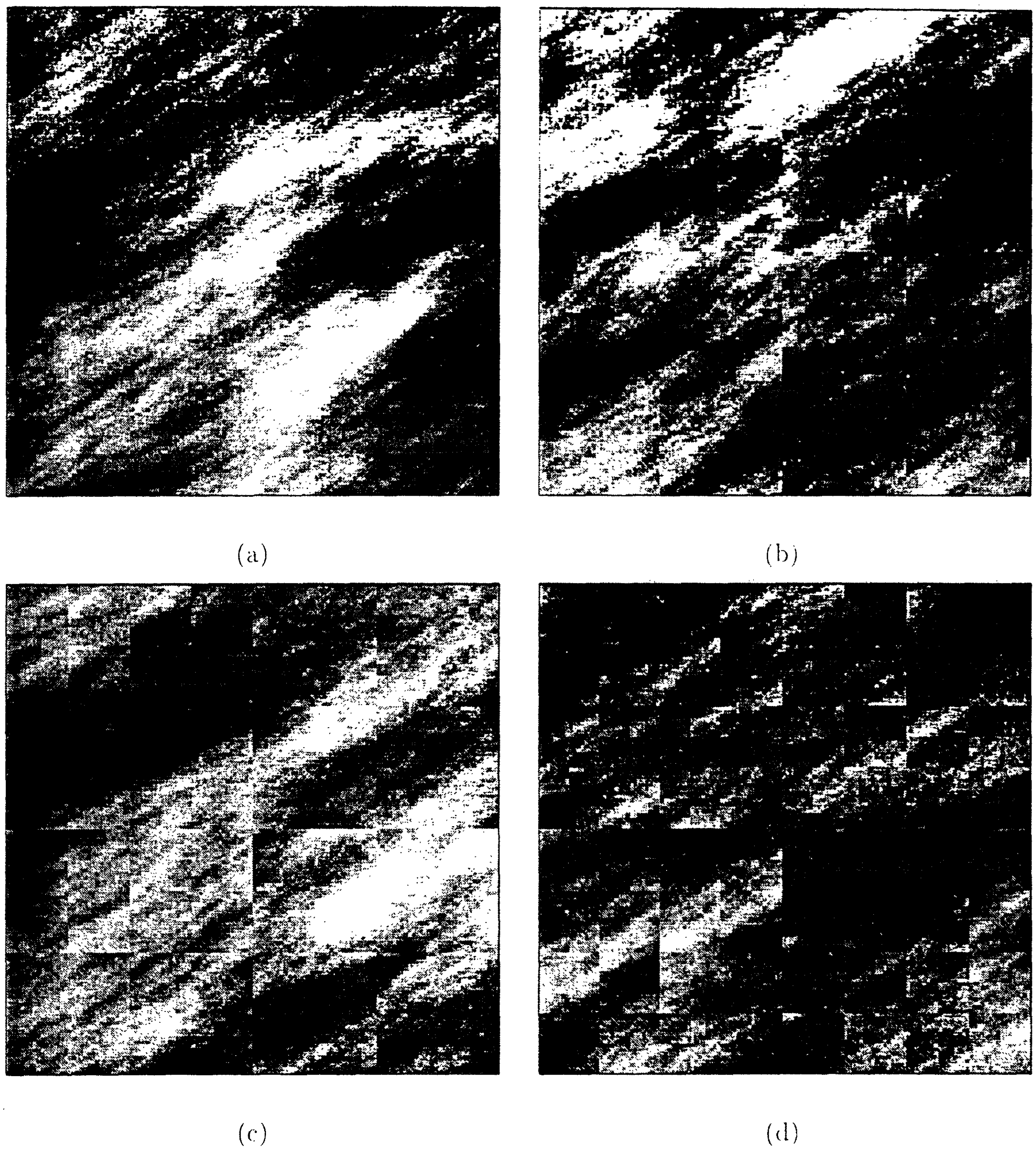
(a) (b)

Figure 7: Comparison of slices of correlation contour plots in the previous figure. (a) A slice along the direction of the major axis of the ellipses in part (a) of the previous figure. (b) A slice along the direction of the minor axis of the ellipses in part (a) of the previous figure. Again, these plots are based on Monte-Carlo simulation, where each point is within 0.005 of its correct value with 95 percent confidence.

While in many applications, these artifacts are devoid of any statistical significance, they may be important in other contexts. One way to eliminate these artifacts is employ a relatively high-order model multiscale model; for instance, as shown in Figure 8a, the 64-th order model is effective in this regard. An alternative, arguably more elegant approach to eliminating these artifacts is to use so-called *overlapping tree* models, in which distinct tree nodes correspond to overlapping portions of the image domain; this idea is described in detail in [9].

6 Conclusions and Suggestions for Future Work

This paper develops elements of a theory for multiscale stochastic realization, focusing on the problem of building multiscale models to realize. either exactly or approximately, pre-specified finest-scale statistics. A key challenge has been to generalize the time-series notion of state vectors serving as an interface between the past and the future of a random process. The generalization is



(a) (b) (c) (d)

Figure 8: These four figures display sample paths of a random field having the correlation function given in (33b, for a 128 × 128 pixel region. The sample paths in (a), (b), (c) and (d) correspond to multiscale models of order 64, 32, 16 and 8, respectively, using Gaussian deviates.

made by introducing a generalized correlation coefficient. which is used to make precise the notion of multiscale state vectors serving to decorrelate multiple subsets of a multiscale process. Once the reduced-order multiscale modeling problem has been formalized, we harness canonical correlation analysis to develop a sub-optimal model-building algorithm. We demonstrate the practicality of this algorithm in problems of random-field estimation and generation. In the context of field generation, we demonstrate an ability to build multiscale processes having a finest-scale correlation matching very closely desired correlations. In the context of field estimation, we build multiscale models that are in turn used to carry out least-squares estimation, with the resulting field estimates having nearly optimal mean-square error.

The work presented raises a number of interesting research questions. What is the optimal solution to the general decorrelation problem addressed in Section 4? Is the bound on state dimension in Proposition 1 tight? In other words, can a multiscale model be devised in which the bound holds with equality at every node? If not, what then constitutes a minimal realization of a given finest-scale covariance? Is the class of internal realizations rich enough to always include a minimal realization? This last question is particularly intriguing, because in the time-series case, the answer is yes [10, 16].

Finally, other interesting questions arise when we consider more carefully the issue of inter-scale propagation of information in multiscale processes. Consider, for example, the problem of building a multiscale model indexed on a second-order tree (i.e., a tree for which $q = 2$) having three scales to realize exactly the following finest-scale covariance:

$$P_{\chi_0} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \qquad (34)$$

One possible exact realization uses the following values for the internal matrices:

$$\begin{aligned} W_0 &= \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix}. \\ W_{0\alpha_1} &= \begin{pmatrix} 0 & 1 \end{pmatrix}. \quad W_{0\alpha_2} = \begin{pmatrix} 1 & 0 \end{pmatrix}. \\ W_{0\alpha_1\alpha_1} &= W_{0\alpha_1\alpha_2} = W_{0\alpha_2\alpha_1} = W_{0\alpha_2\alpha_2} = 1. \end{aligned} \tag{35}$$

A valid alternative, which also leads to an exact realization, is to replace W_0 in (35) with W_0',

$$W_0' = \operatorname{diag}(1, 1, 1, 1),$$

while retaining the same values as in (35) for the other internal matrices. There is an important difference between the models that result from these two choices for the internal matrices. The first choice leads to a model in which coarse-scale information is preserved in its journey to the finest scale; in particular, we see that $x(s) = W_s\xi_s$. and so indeed the first realization is internal. On the other hand, the second choice leads to a model in which information is lost in its journey to the finest scale. In fact, by using W_0' in lieu of W_0, we have somewhat perversely created a multiscale model in which the *entire* finest-scale process is generated at the root node, and then some of this information is immediately discarded in the transition to the middle scale, whence new values for this discarded information are generated in the transition to the third, finest scale. Although the finest scale process does have the correct, desired correlation, the realization is *not* internal[5]. In particular, $x(0) \neq W_0\xi_0$.

In light of this example, how can our modeling approach be refined to propagate information more explicitly from scale to scale? One answer is presented in [9], though the suggested approach is not numerically practical. As a related issue, how can we extend our modeling approach to realize jointly pre-specified fine *and* coarse scale statistics?

All of these unanswered questions highlight the fact that multiscale stochastic realization is a relatively new subject, with this paper serving only as a beginning. Many interesting challenges remain.

[5] This example demonstrates that condition (13) is necessary but not sufficient for an exact, *internal* realization.

A Proof of Proposition 2

In this appendix, we complete the proof of Proposition 2. by establishing the validity of (18a) and (18b). For this purpose, we continue to use the notation established in Section 4.1.

We begin by making explicit the connection between the value of $\bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1)$ and the cross-covariance D between μ_1 and μ_2. To proceed, we define the linear least-squares residual $\tilde{\mu}$ via

$$\tilde{\mu} \equiv \begin{pmatrix} \tilde{\mu}_1^T & \tilde{\mu}_2^T \end{pmatrix}^T \equiv \mu - E(\mu \mid W\mu), \tag{36}$$

where by elementary theory of linear least-squares estimation theory,

$$P_{\tilde{\mu}} = P_\mu - P_\mu W^T \left(W P_\mu W^T\right)^{-1} W P_\mu \tag{37}$$

For this analysis, we set $W = (W_1\ 0)$, so that $W\mu = W_1\mu_1$. In terms of (37), we then define the sets F_1 and F_2 as follows:

$$F_i \equiv \left\{f;\ f^T P_{\tilde{\mu}_i} f = 1\right\} \qquad (i = 1, 2). \tag{38}$$

Now, using the definitions in Section 2.3, we find that if either F_1 or F_2 is empty, then $\bar{\rho}(\mu_1, \mu_2 | W_1\mu_1) = 0$, and otherwise

$$\bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1) = \max_{f_1 \in F_1, f_2 \in F_2} f_1^T P_{\tilde{\mu}_1\tilde{\mu}_2} f_2. \tag{39}$$

In general, the sub-matrices $P_{\tilde{\mu}_1}$, $P_{\tilde{\mu}_2}$ and $P_{\tilde{\mu}_1\tilde{\mu}_2}$ of $P_{\tilde{\mu}}$ in (37) can have messy analytical forms. However, if W_1 has orthonormal rows (i.e., $W_1 \in \mathcal{N}_\lambda$, for some λ), then simplification is possible. We begin by noting that

$$\bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1) = \bar{\rho}(\mu_1', \mu_2 \mid W_1'\mu_1') \tag{40}$$

where μ_1' and W_1' are related respectively to μ_1 and W_1 via unitary (but otherwise arbitrary) matrix U:

$$\mu_1' \equiv U^T\mu_1, \quad W_1' \equiv W_1 U, \quad \tilde{\mu}'_1 \equiv \mu_1' - E(\mu_1' \mid W_1'\mu_1') \tag{41}$$

Letting $W_1^\perp$ be a matrix whose rows form an orthonormal basis for the nullspace of W_1, we set $U = (W_1^T \;\; (W_1^\perp)^T)$, so that $W_1' = (I_\lambda \;\; \mathbf{0})$. Then, thanks to (37),

$$\begin{pmatrix} P_{\bar{\mu}'_1} & P_{\bar{\mu}'_1\bar{\mu}_2} \\ P^T_{\bar{\mu}'\mu_2} & P_{\bar{\mu}_2} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{m_1-\lambda} \end{pmatrix} & \begin{pmatrix} \mathbf{0} \\ W_1^\perp D \end{pmatrix} \\ \begin{pmatrix} \mathbf{0} \\ W_1^\perp D \end{pmatrix}^T & I_{m_2} - D^T W_1^T W_1 D \end{pmatrix}. \tag{42}$$

Finally, adapting (40)-(42) to (36)-(38), the following lemma results:

Lemma 2 *Let $W_1 \in \mathcal{N}_\lambda$, $0 \leq \lambda < m_1$, and let $W_1^\perp$ be a matrix whose rows form an orthonormal basis for the nullspace of W_1. Then,*

$$\bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1) = \max_{g_1 \in G_1,\, g_2 \in G_2} \left\{ g_1^T W_1^\perp D g_2 \right\} \tag{43a}$$

$$= \max_{g_2 \in G_2} \| W_1^\perp D g_2 \|_2, \tag{43b}$$

where G_1 and G_2 denote the following sets:

$$\begin{aligned} G_1 &= \left\{ g \in \mathcal{R}^{m_1-\lambda};\; g^T g = 1 \right\}, \\ G_2 &= \left\{ g \in \mathcal{R}^{m_2};\; g^T (I_{m_2} - D^T W_1^T W_1 D) g = 1 \right\}. \end{aligned}$$

As a direct consequence of this lemma,

$$\bar{\rho}(\mu_1, \mu_2 \mid (I_\lambda \;\; \mathbf{0})\,\mu_1) = \begin{cases} d_{\lambda+1}, & 0 \leq \lambda < m_{12} \\ 0 & \lambda \geq m_{12} \end{cases} \tag{44}$$

This fact establishes (18b).

What remains is to establish (18a). To proceed, we temporarily constrain W to have either of the two forms

$$W = \begin{pmatrix} W_1 & \mathbf{0} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \mathbf{0} & W_2 \end{pmatrix}, \tag{45}$$

and we find a matrix $W \in \mathcal{N}_\lambda$ that minimizes $\bar{\rho}(\mu_1, \mu_2 \mid W\mu)$, subject to this additional constraint. The following lemma summarizes the key result here.

Lemma 3

$$\begin{aligned}\min_{W_1 \in \mathcal{N}_\lambda} \bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1) &= \bar{\rho}(\mu_1, \mu_2 \mid \begin{pmatrix} I_\lambda & \mathbf{0} \end{pmatrix} \mu_1) = \bar{\rho}(\mu_1, \mu_2 \mid \begin{pmatrix} I_\lambda & \mathbf{0} \end{pmatrix} \mu_2) \\ &= \min_{W_2 \in \mathcal{N}_\lambda} \bar{\rho}(\mu_1, \mu_2 \mid W_2\mu_2) = d_{\lambda+1}.\end{aligned}$$

Proof: Thanks to (44), it is sufficient to show that for all $W_1 \in \mathcal{N}_\lambda$,

$$\bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1) \;\geq\; d_{\lambda+1}. \tag{46}$$

To establish (46), we fix $W_1 \in \mathcal{N}_\lambda$. Referring back to Lemma 2, and in particular (43a), we devise particular values for $g_1 \in G_1$ and $g_2 \in G_2$ for which

$$g_1^T(W_1^\perp D)g_2 \;\geq\; d_{\lambda+1}, \tag{47}$$

thus implying that $\bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1) \;\geq\; d_{\lambda+1}$.

To establish (47), we first note that at least one of the unit vectors $e_1^T, e_2^T, \ldots, e_{\lambda+1}^T$ must belong to the row space of $W_1^\perp$, which itself has a dimension of $m_1 - \lambda$; let us suppose that e_j^T belongs, with $h^T W_1^\perp = e_j^T$ for some $h \in \mathcal{R}^{m_1 - \lambda}$. Now, exploiting the orthonormality of the rows of $W_1^\perp$, we see that $h \in G_1$, and hence we let $g_1 = h$. Also, we let $g_2 = e_j$, where the fact that $De_j = d_j e_j = d_j(W_1^\perp)^T g_1$, implies that $W_1 De_j \;=\; \mathbf{0}$, so that indeed $e_j \in G_2$. But for these values for g_1 and g_2, $g_1^T(W_1^\perp D)g_2 = d_j \geq d_{\lambda+1}$, thus establishing (47) and completing the proof. **QED**.

Next, we establish that in fact there is no loss of optimality in the additional constraint in (45). The following lemma summarizes the key result.

Lemma 4 *For any matrix $W \in \mathcal{N}_\lambda$ there exists a pair of matrices $W_1 \in \mathcal{M}_{\lambda_1}$, and $W_2 \in \mathcal{M}_{\lambda_2}$, with $\lambda_1 + \lambda_2 \leq \lambda$, such that $\bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1, W_2\mu_2) \leq \bar{\rho}(\mu_1, \mu_2 \mid W\mu)$.*

Proof: We begin by expressing the matrix W in terms of its constituent rows as

$$W \;=\; \begin{pmatrix} W_1 & W_2 & \cdots & W_\lambda \end{pmatrix}^T,$$

where the column vector W_i, for $i = 1, 2, \ldots, \lambda$, can itself be decomposed as

$$W_i \;=\; \begin{pmatrix} W_{i,1} \\ W_{i,2} \end{pmatrix}, \qquad W_{i,j} \in \mathcal{R}^{m_j},\; j = 1, 2.$$

Let us suppose that for some particular i, say i_1, $W_{i_1,1} \neq \mathbf{0}$, and $W_{i_1,2} \neq \mathbf{0}$. We demonstrate that W_{i_1} can be replaced with one of the two vectors V_1 or V_2, where

$$V_1 \equiv \begin{pmatrix} W_{i_1,1} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad V_2 \equiv \begin{pmatrix} \mathbf{0} \\ W_{i_1,2} \end{pmatrix}.$$

with no incurred increase in the value of $\bar{\rho}(\mu_1, \mu_2 \mid W\mu)$, after the replacement.

We now define $\tilde{\mu}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$ and $P_{\tilde{\mu}}$ as in (36) and (37). From (37), it immediately follows that

$$P_{\tilde{\mu}} W^T = \mathbf{0}, \tag{48}$$

which in turn implies that $P_{\tilde{\mu}} W_i = \mathbf{0}$, so that

$$W_{i_1,1}^T P_{\tilde{\mu}_1} W_{i_1,1} = W_{i_1,2}^T P_{\tilde{\mu}_2} W_{i_1,2} = -W_{i_1,1}^T P_{\tilde{\mu}_1,\tilde{\mu}_2} W_{i_1,2}.$$

There are now two possibilities: either $W_{i_1,1}^T P_{\tilde{\mu}_1} W_{i_1,1} > 0$, or $W_{i_1,1}^T P_{\tilde{\mu}_1} W_{i_1,1} = 0$. The first implies that $\bar{\rho}(\mu_1, \mu_2 \mid W\mu) = 1$, in which case there can certainly be no harm in our replacement strategy. The second implies that $P_{\tilde{\mu}_1} V_1 = \mathbf{0}$, and $P_{\tilde{\mu}_2} V_2 = \mathbf{0}$, which, in turn, means that there exist unique vectors h_1 and h_2 in $\mathcal{R}^\lambda$ such that $V_i = W^T h_i$, $(i = 1, 2)$. Now, by exhaustively considering the possibilities, one can verify that at least one of h_1 and h_2 must have a non-zero value in its i_1-th component, for otherwise, W could not have full row rank. If h_1 (h_2) has this property, then we can replace W_{i_1} with V_1 (V_2) with no change in the value of $\bar{\rho}(\mu_1, \mu_2 \mid W\mu)$. **QED.**

To make clear the consequences of this lemma, let us suppose that $W^* \in \mathcal{N}_\lambda$ minimizes $\bar{\rho}(\mu_1, \mu_2 \mid W\mu)$. Then, from the lemma we know there exist matrices $W_1^* \in \mathcal{M}_{\lambda_1}$ and $W_2^* \in \mathcal{M}_{\lambda_2}$ (for some λ_1 and λ_2 such that $\lambda_1 + \lambda_2 \leq \lambda$) such that $\bar{\rho}(\mu_1, \mu_2 \mid W^*\mu) = \bar{\rho}(\mu_1, \mu_2 \mid W_1^*\mu_1, W_2^*\mu_2)$. Then, fixing W_1^*, let us define $\tilde{\mu}_i \equiv \eta_i - E(\mu_i \mid W_1^*\mu_1)$, $(i = 1, 2)$, which we use to see that

$$\begin{aligned}
\bar{\rho}(\mu_1, \mu_2 \mid W^*\mu) &= \min_{W_2 \in \mathcal{M}_{\lambda_2}} \bar{\rho}(\mu_1, \mu_2 \mid W_1^*\mu_1, W_2\mu_2) = \min_{W_2 \in \mathcal{M}_{\lambda_2}} \bar{\rho}(\tilde{\mu}_1, \tilde{\mu}_2 \mid W_2\tilde{\mu}_2) \\
&= \min_{\bar{W}_1 \in \mathcal{M}_{\lambda_2}} \bar{\rho}(\tilde{\mu}_1, \tilde{\mu}_2 \mid \bar{W}_1\tilde{\mu}_1) = \min_{\bar{W}_1 \in \mathcal{M}_{\lambda_2}} \bar{\rho}(\mu_1, \mu_2 \mid W_1^*\mu_1, \bar{W}_1\mu_1) \\
&= \min_{W_1 \in \mathcal{M}_\lambda} \bar{\rho}(\mu_1, \mu_2 \mid W_1\mu_1).
\end{aligned}$$

The first equality follows directly from Lemma 3. The second follows from the definition of $\tilde{\mu}_i$, while the third line follows from Lemma 2. The fourth equality follows again from the relation between

$\tilde{\mu}_i$ and μ_i, and finally, the fifth follows from the fact that both of the vectors of conditioning information in the fourth line of functions only of μ_1. The proof of Proposition 2 is now complete. **QED**.

B Proof of Proposition 3

In this appendix, we complete the proof of Proposition 3, by establishing the validity of (25). For this purpose, we continue to use the notation established in Section 4.1.

We begin by fixing W_1, which we assume without loss of generality to have orthonormal rows. We know from (44) that $\bar{\rho}(\mu_1, \mu_2) = d_1$. Combining this fact with (43b), it follows that the Proposition will be proved if we can show that

$$\max_{g_2 \in G_2} \| W_1^{\perp} D g_2 \|_2^2 \;\le\; d_1^2. \tag{49}$$

To establish (49), we first note that since the rows of $W_1^{\perp}$ form an orthonormal basis for the null space of W_1, we have that $\forall x$,

$$\| x \|_2^2 \;=\; \| W_1 x \|_2^2 + \| W_1^{\perp} x \|_2^2 \,. \tag{50}$$

Since $\forall g_2 \in G_2$,

$$g_2^T (I - D^T W_1^T W_1 D) g_2 \;=\; \| g_2 \|_2^2 - \| W_1 D g_2 \|_2^2 = 1, \tag{51}$$

we can apply (50) in (51) with $x = D g_2$ to see that $\forall g_2 \in G_2$,

$$\| W_1^{\perp} D g_2 \|_2^2 \;=\; \| D g_2 \|_2^2 - \| g_2 \|_2^2 + 1, \quad \forall g_2 \in G_2. \tag{52}$$

But

$$\begin{aligned}
\min_{g_2 \in G_2} \left\{ \| g_2 \|_2^2 - \| D g_2 \|_2^2 \right\} &= \min_{g_2 \neq 0} \left\{ \frac{g_2^T \left(I - D^T D\right) g_2}{g_2^T \left(I - D^T W_1^T W_1 D\right) g_2} \right\} \\
&\ge \min_{g_2 \neq 0} \left\{ \frac{g_2^T (I - D^T D) g_2}{g_2^T g_2} \right\} \min_{g_2 \neq 0} \left\{ \frac{g_2^T g_2}{g_2^T (I - D^T W_1^T W_1 D) g_2} \right\} \\
&= \mathrm{eig}_{min} \left(I - D^T D\right) \, \mathrm{eig}_{min} \left[\left(I - D^T W_1^T W_1 D\right)^{-1} \right] \\
&= (1 - d_1^2)\,(1)
\end{aligned} \tag{53}$$

where $\text{eig}_{min}(\cdot)$ denotes the smallest eigenvalue of the enclosed matrix expression. In the third line, we have used Rayleigh's principle [18]. which asserts that for any pair of symmetric. positive definite matrices A and B.

$$\min_{x \neq 0} \frac{x^T B x}{x^T A x} = \text{eig}_{min}(A^{-1}B).$$

By combining (52) and (53), the desired result (49) is established. **QED**.

References

[1] H. Akaike. "Markovian Representation of Stochastic Processes by Canonical Variables." In *SIAM Journal of Control*, vol. 13, no. 1. January, 1975.

[2] H. Akaike. "Stochastic Theory of Minimal Realization." In *IEEE Transactions on Automatic Control*, vol. 19, no. 6, December, 1974.

[3] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Inc., New York. 1958.

[4] K.C. Chou, A.S. Willsky, and A. Benveniste. "Multiscale recursive estimation, data fusion and regularization." In *IEEE transactions on Automatic Control*, Vol. 39, No. 3, March, 1994.

[5] H. Derin and P.A. Kelly, "Discrete-Index Markov-type Random Processes." In *Proceedings of the IEEE*, Vol. 77, No. 10, October, 1989.

[6] U.B. Desai and D. Pal. "A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realizations." In *Proceedings of the 21st Conference on Decision and Control*, pp. 1105-1114, 1982.

[7] P. Fieguth, A. Willsky, W. Karl, C. Wunsch, "Multiresolution Optimal Interpolation and Statistical Analysis of TOPEX/POSEIDON Satellite Altimetry," *IEEE Trans. Geoscience and Remote Sensing* (33) #2. pp.280–292, March 1995

[8] H. Hotelling. "Relations between two sets of variates." In *Biometrika*. Vol. 28, pp. 321-377. 1936.

[9] W. Irving. *Multiresolution Stochastic Realization and Model Identification with Applications to Large-Scale Estimation Problems* PhD thesis, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, September, 1995.

[10] A. Lindquist and G. Picci. "On the Stochastic Realization Problem." In *SIAM Journal of Control and Optimization*, Vol. 17, No. 3, May, 1979.

[11] M. Luettgen and A.S. Willsky, "Likelihood calculations for a class of multiscale stochastic models, with applications to texture discrimination." Technical Report LIDS-P-2115, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 1993.

[12] M. Luettgen, W.C. Karl, A.S. Willsky, and R.R. Tenney, "Multiscale representations of Markov random fields." In *IEEE Transactions on Signal Processing*, December, 1993.

[13] M. Luettgen, W. Karl, A. Willsky, "Efficient Multiscale Regularization with Applications to the Computation of Optical Flow." In *IEEE Transactions on Image Processing*, vol 3, No. 1, pp. 41-64, 1994.

[14] D.F. Morrison. *Multivariate Statistical Methods*. McGraw-Hill Book Company, New York. 1967.

[15] R.J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., New York. 1982.

[16] G. Picci. *Stochastic Realization of Gaussian Processes*. In *Proceedings of the IEEE*, Vol. 64, No. 1, January, 1974.

[17] M.A. Sironvalle, "The Random Coin Method: Solution of the Problem of Simulation of a Random Function in the Plane." In *Mathematical Geology*, vol 12, no. 1, 1980.

[18] G. Strang, *Linear Algebra and its Applications*. Harcourt Brace Jovanovich, Publishers, San Diego, 1988.